Location and Extraction of Broadcast in News Video Based on QGMM and BIC

Ling Guo [1] Ying-Chun Shi [1,2] Xian-Zhong Zhou [1] Feng Zhang [1]
1.Dept. of Automation, Nanjing University of Science & Technology, NanJing 210094
2.Simulation Center, WuHan Institute of Communication Command, WuHan 430010
E-mail:laura0955@163.com yingchun1004@163.com

Abstract

An algorithm on location and extraction of broadcast in news video is proposed in this paper. Firstly, input audio stream is divided into speech and non- speech segments by VQ (Vector Quantification) after a set of new features representing audio segments' time-variant characteristics are extracted, including HZCRR (High Zero-crossing Rate Ratio), LSTER (Low Short-time Energy Ratio) and HBFERR (High Basic-frequency-energy Rate Ratio), etc. Then a QGMM (Quasi Gaussian Mixture Model) is presented to describe the speaker's identity and BIC (Bayesian Information Criterion) is used to detect speaker change. Finally speaker clustering is carried out with BIC, and location and extraction of broadcast is realized based on rules. Satisfactory results from experiments prove the effectiveness of this algorithm.

1. Introduction

With the development of computer technology and Internet, the amount of video information available is increasing dramatically. The problem for users has switched from shortage of information to quick and reasonable retrieval of required information from large quantities of information. The technology of CBVR (Content-based Video Retrieval) appeared at the early 20th century to solve this problem. Nowadays, some prototype systems have been developed, such as VideoQ [1], WebSeek [2], Informedia [3], etc. They describe video contents with only low-level video features, such as color, textures, shape, motion, time-series, time-span, which usually appear in the form of statistical data. But people prefer advanced semantical features, which are hard to extract from visual low-level features, because video is a spatio-temporal combination of text, image, audio and other multi-modal information, and each mode enriches semantical information. In order to overcome the drawbacks of video representation by mere visual semantics, researchers begin to try combining audio (speech, music), text (caption, title), and other information in the retrieval and abstraction of video [6-9].

We show special interests in news video, in which important information, such as big events, the time, place and characters concerned, is broadcasted, providing highly abstracted semantics about the video. So a correct location, extraction and recognition of a video can make browse and retrieval more convenient by advanced-semantics-based segmentation and labeling. Thus, we present an algorithm on automatic location and extraction of broadcast in news video based on QGMM and BIC. The flow diagram of the algorithm is shown in Fig. 1. In the rest of the paper, the analysis of features and algorithm is presented in

Proceedings of the 2005 The Fifth International Conference on Computer and Information Technology (CIT'05)
0-7695-2432-X/05 $20.00 © 2005 IEEE

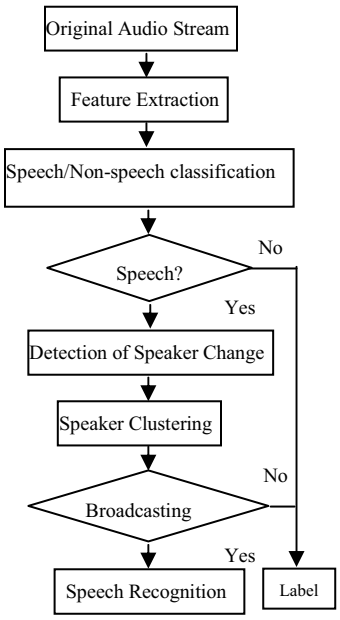

Figure1. Algorithm diagram

section 2 and section 3 respectively and details about experiments and results are given in section 4.

2. Feature analysis

In papers about audio content analysis, static statistical features such as short-time energy, zero-crossing rate, central frequency, etc., are widely used to describe audio signals, leaving dynamic features ignored. We use HZCRR, LSTER and HBFERR to depict time-variant statistical features of ZCR (Zero-crossing rate), STE(short-time energy) and BFER (Basic-Frequency-Energy Rate) of the audio, and MFCC (Mel-Frequency Cepstral Coefficients) to describe speaker's identity.

HZCRR is defined as the ratio of the number of frames with a ZCR above1.5 times of average zero-crossing rate in an audio segment:

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(ZCR(n) - 1.5avZCR) + 1]$$, where n is the frame index, ZCR(n) is the zero-crossing rate at the nth frame, N is the total number of frames, avZCR is the average ZCR in an audio segment, and sgn[.] is a sign function. Unlike music, speech signals usually consist of alternative sonant and surd, so their HZCRR is larger than that of music. Figure2 shows the contrast.

LSTER is defined as the ratio of the number of frames with a STE less than 0.5 time of average short-time energy in an audio segment:

$$LSTER = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(0.5avSTE - STE(n)) + 1]$$, where N is the total number of frames, STE(n) is the short-time energy at the nth frame, and avSTE is the average STE in an audio segment. Because speech usually has much more silent frames than music, LSTER of speech is larger than that of music, as Fig.3 shows.

HBFERR is defined as the ratio of the number of frames with a BFER more than 1.5 time of average basic-frequency-energy ratio in an audio segment:

$$HBFERR = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(BFER(n) - 1.5avBFER) + 1]$$, where BFER is defined as the ratio of energy of frequency between 0~1.5kHz to the energy of all frequencies:

$$BFER = \sum_{n=1}^{N-1}\sum_{k=1}^{1500}[A(n,k)]^2 \Big/ \sum_{n=1}^{N-1}\sum_{k=1}^{K}[A(n,k)]^2$$, where $A(n,k)$ is Fourier transform of an audio frame. The energy of speech stream mainly centralizes between 0 and 1.5kHz，while that of other types of audio streams either distributes widely or centralizes at high-frequency part, so speech usually has a larger HBFFR than other types of audio streams, as Fig. 4 shows.

MFCC and its covariance matrix can represent speaker's identity, with the former describing the difference in speech organs (inborn), and the latter describing the difference in organs' actions when pronouncing (acquired) [11]. MFCC is results of filtering Fourier Transform frequency coefficients by triangular filters group in Mel scale frequency domain.

$$C_n = \sqrt{\frac{2}{K}\sum_{k=1}^{K}(\log S_k)\cos[n(k-0.5)\pi / K]}$$, where S_k is the output of triangular filter and K is the number of filters.

3. Algorithm analysis

3.1 Speech and non-speech classification

After feature analysis, a VQ algorithm[10] based on HZCRR, LSTER, SF and BFFER is applied to classification of speech and non-speech. Here, all audio streams are sampled into 22.050KHz, 16bit data, mono-channel data, and then pre- emphasized and

Proceedings of the 2005 The Fifth International Conference on Computer and Information Technology (CIT'05)
0-7695-2432-X/05 $20.00 © 2005 IEEE

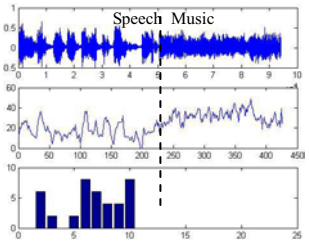

Figure 2. Original audio stream, ZCR, HZCRR

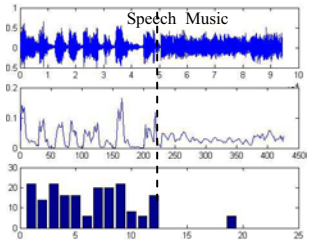

Figure 3. Original audio stream, STE, LSTER

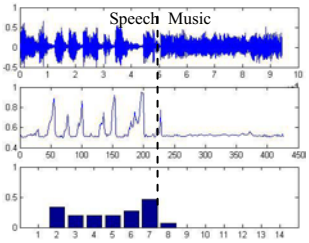

Figure 4. Original audio stream, BFER, HBFERR

segmented into 2s sub-segments with 1s overlapping, which are further divided into non-overlapping 20ms frames. Then features such as HZCRR, LSTER, HBFERR, are extracted and VQ algorithm is applied to codebooks training and audio segments classification.

The four training steps include: extraction of feature vectors from training audio to compose a feature vector set, generation of a code book by LBG, optimization of the code book by repeated training and storage of the code book.

The three classification steps include: ①extraction of feature vector serials: $X_1, X_2, \ldots, X_M$, from audio;② quantification of each vector by trained code book and computation of each mean quantification error by $D_i = \frac{1}{M}\sum_{n=1}^{M}\min_{1\le j\le L}[d(X_n, Y_j^i)]$, $j = 1,2,\cdots,L, i = 1,2,\cdots,N$, where Y_l^i is the jth vector in ith codebook, and $d(X_n, Y_j^i)$ is the distance between classification vector X_n and code vector Y_j^i;③selection of a class with the least mean quantification error as the classification result.

3.2 Speaker change detection

Firstly we remove non-speech frames from the audio stream to get a speech stream with the results of speech and non-speech classification, segment it into non-overlapping 1s audio segments and divide them into non-overlapping 20ms frames further, and extract each frame's MFCC feature and calculate MFCC and MFCC covariance matrix features.

Secondly, speaker's identity is modeled after feature extraction to detect speaker change. Modeling from a single segment often results in a biased estimation of speaker models due to the shortage of data, so its' essential to get as much data as possible to correct the model. We use a GMM to depict speaker models. Probability density of classical M-order GMM is a weighted sum of M Gaussian probability density. An improved algorithm to estimate GMM is proposed here and it works well in speaker clustering, although a lower calculating complexity and less memory or disk storage are achieved at the cost of accuracy.

Suppose the current speaker model $G_i \sim N(\mu, C)$ is obtained from the previous (M-1) sub-segments and there's no potential speaker change between the (M-1)th and the Mth speech segment, which means the two segments belong to the same speaker. Then we update the current speaker model G_i with the feature data of the Mth segment. Suppose the model of Mth speech segment is represented by $N(\mu_M, C_M)$, according to statistical theory, then $N(\mu', C')$ could be derived from the following method: $\mu' = \frac{n\mu + n_M\mu_M}{n + n_M}$,

$$C' = \frac{n}{n+n_M}C + \frac{n_M}{n+n_M}C_M + \frac{n\cdot n_M}{(n+n_M)^2}\cdot(\mu-\mu_M)(\mu-\mu_M)^T \quad ,$$

where n and n_M are the number of feature vectors of the model $N(\mu, C)$ and $N(\mu_M, C_M)$ respectively. The last term of the formula is determined by the means easily affected by environment conditions [11]. Thus, in practice, we ignore it to compensate the effect of environment conditions. Then the above formula is

Proceedings of the 2005 The Fifth International Conference on Computer and Information Technology (CIT'05)
0-7695-2432-X/05 $20.00 © 2005 IEEE

simplified as: $\mu' = \frac{n\mu + n_M \mu_M}{n + n_M} = \sum_{i=1}^{M} n_i \mu_i \Big/ \sum_{i=1}^{M} n_i$,

$C' = \frac{n}{n+n_M} C + \frac{n_M}{n+n_M} C_M = \sum_{i=1}^{M} n_i C_i \Big/ \sum_{i=1}^{M} n_i$, which is the QGMM[9] that we use to model speakers.

Finally, we use BIC [12] to detect speaker change. Let $x = \{x_i : i = 1, \cdots, N\}$ be the data set we are modeling; let $M = \{M_i : j = 1, \cdots, K\}$ be the candidates of desired parametric models, compute the maximum likelihood function for each model M separately, [$L(x, M)$], and denote $\#(M)$ as the number of parameters in the model M. Then BIC criterion is defined as: $BIC(M) = \log L(x, M) - \lambda \frac{1}{2} \#(M) \times \log(N)$, where the penalty weight $\lambda = 1$. The BIC procedure is to choose the model that maximizes BIC, which can be derived from a large-sample version of Bayesian procedures in the case of independent, identically distributed observation and linear models. Denote $X = \{x_i \in R^d, i = 1, \cdots, N\}$ as the series of MFCC + △ MFCC vectors extracted from the audio stream and assume X obeys an independent multivariate Gaussian process: $x_i \sim N(\mu_i, C_i)$, where μ_i is the mean vector and C_i is the full covariance matrix. Assume that a speaker change occurs at time I, and C, C_1, C_2 are the sample covariance matrices from $\{x_1, \cdots, x_N\}, \{x_1, \cdots, x_i\}, \{x_{i+1}, \cdots, x_N\}$ respectively. Then BIC can be expressed as $BIC(i) = R(i) - \lambda P$, where $R(i) = n \log|C| - n_1 \log|C_1| - n_2 \log|C_2|$ is the likelihood function item, $P = \frac{1}{2}(d + \frac{1}{2} d(d+1)) \log n$ is the penalty item, $\lambda = 1$, d is the dimension of the space, and the number of parameters is $d + \frac{1}{2} d(d+1)$. $\max_i BIC(i) > 0$ indicates that the two speech segments should be described by two different Gaussian models: $N(\mu_1, C_1)$ and $N(\mu_2, C_2)$, that is, a speaker change happens and the change point can be expressed as $\hat{t} = \arg\max_i BIC(i)$, as Fig.5 illustrates; otherwise, the two speech segments should be described by one Gaussian model: $N(\mu, C)$ and no speaker change happens. The speaker's QGMM is corrected by the second speech segment.

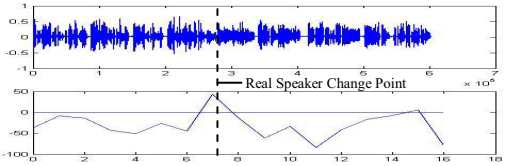

Figure 5. Original audio atream, BIC value

3.3 Location and extraction of broadcast

After the detection of speaker change, speaker clustering is needed to identify broadcaster's speech using rules, which is carried out by the following steps:

Step1: sort the divided speech segments by time-standing length in descending order;

Step2: select the longest one from the non-clustering speech segments as the classification seed, then use BIC criterion for clustering. Calculate $BIC(C_1, C_2)$ between classification seed's C_1 and all the other non-clustering speech segments' C_2 by:

$$BIC(c_1, c_2) = (n_1 + n_2) \log|C| - n_1 \log|C_1| - n_2 \log|C_2|$$
$$- \lambda \frac{1}{2} (d + \frac{1}{2} d(d+1)) \log(n_1 + n_2)$$

$BIC(C_1, C_2) <= 0$ indicates that they belong to the same speaker, so mark it; otherwise consider it as non-clustering segments;

Step3: If there are still some non-clustering segments, go to step 2; otherwise, end.

Then location and extraction of broadcast is realized by sorting clustering results by classifications in descending order firstly, then defining n, the number of anchors, with priori knowledge and choosing first n classifications of speech as broadcast segments, and finally recognizing broadcast segments using Microsoft's Speech SDK.

Proceedings of the 2005 The Fifth International Conference on Computer and Information Technology (CIT'05)
0-7695-2432-X/05 $20.00 © 2005 IEEE

4. Experiment and Results

We have chosen four pieces of CCTV-1 news broadcast for test. The evaluation of algorithm performance is described with recall and precision, which are defined as $\mathrm{Re}call = Detected/Original$ and $\mathrm{Pr}ecision = Detected\,correctly/Original$, respectively. For detection of speech and non-speech, we use two pieces of the news broadcast as training data for VQ training, and the rest two for test. For detection of speaker change and broadcast, we use all the news broadcast and the number of speakers is 3. Detailed results are listed in Table1.

Table 1. Experiment results

	Original	Detected	False	Miss	Recall	Precision
Speech segment	3441	3342	96	195	94.3%	97.1%
Speaker change	237	261	49	25	89.7%	81.3%
Broadcast segment	5578	5861	1201	918	83.6%	79.5%

5. Conclusions

In the broadcast of news video, all the important information is reported, providing highly-abstracted semantics and making advanced-semantics-based video browse and retrieval possible. In this paper, we present an algorithm on location and extraction of broadcast, in which audio is firstly segmented into speech and non-speech segments by VQ; then BIC is used to detect the change of speakers, a QGMM model is presented to describe the speaker's model, and speaker clustering is carried out with BIC; finally location and extraction of broadcast is realized. New features are also proposed to describe audio segments' time-variant characteristics, including HZCRR, LSTER, HBFERR. Experiments prove the effectiveness of this algorithm. Our future research will focus on how to select audio features to achieve a better representation of speaker's identity and how to select speaker's model and judgmental rules to achieve higher detection accuracy.

References:

[1] Shi-Fu Chang et al, "videoQ: An Automated Content Based Video Search System Using Visual Cues", *ACM Multimedia*, Seattle Washington USA, 1997: 33-34

[2] http://www.ctr.Columbia.edu/webseek

[3] Wactlar H.D. et al, "Intelligent Access to Digital Video: Informedia Project", *IEEE Computer*, 1999, 29(6):46-52

[6] Lie Lu et al, "Content Analysis for Audio Classification and Segmentation", *IEEE Trans on speech and audio processing*, 2002, 10(7):504-516

[7] Wu Fei et al, "Compressed Feature Based TV Program Classification and Retrieval Using Speaker Identification", *PR&AR 2002*, 15(1):21-26

[8] Zhuang Yueting et al, "Hidden Markov Model Based Broadcast News Segmentation and Classification", *Journal of computer research and development*, 2002, 39(9):1057-1063

[9] Zhuang Yueting et al, "Automatic Caption Location and Extraction in Digital Video Based on Support Vector Machine", *Journal of computer aided design and computer graphics*, 2002, 14(8):750-753

[10] Zhao Li, *Speech signal processing*, china machine press, 2003.3

[11] J.P.Campbell, "Speaker Recognition: A tutorial", *Proc. IEEE*, 1997, 85(9): 1437-1462.

[12] S.S.Chen et al, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion", *Proc. DARPA broadcast news transcription and understanding workshop*, 1998: 69-72.